*This is the main submission document. **Save and rename this document filename with your <u>registered full name</u> as Prefix before submission.** Submit both this word document and PDF format.*

Class	1
Full Name	Riley Ang Xile
University Email	rile0001@e.ntu.edu.sg
Matriculation Number	U2022075E

** : Delete and replace as appropriate.*

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

Please insert an "X" within the square bracket below to indicate your selection.

[X] I have read and accept the above.

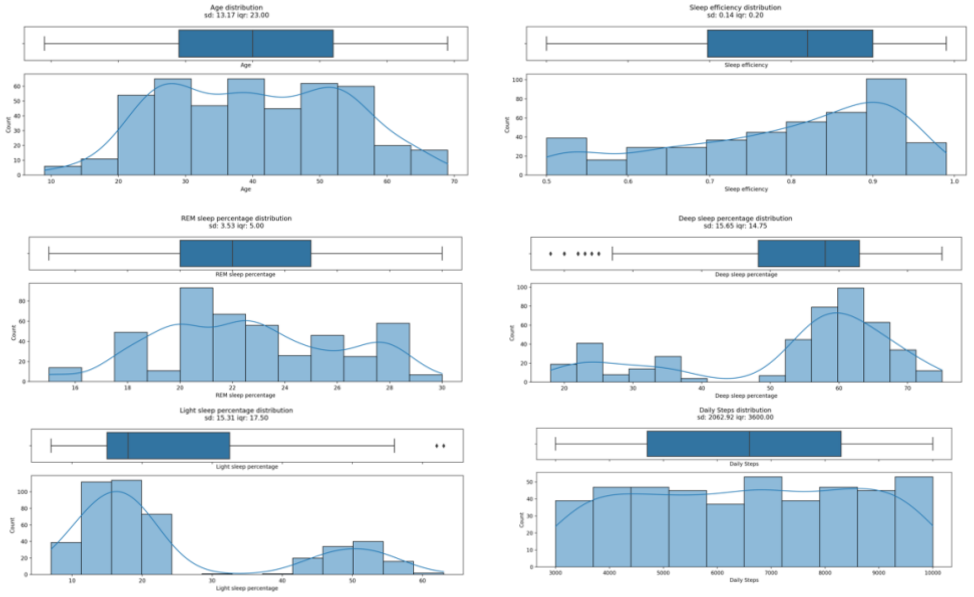
Table of Contents

Answer to Q1:

We are presented with a dataset filled with different metrics that measured one's sleep including sleep duration, efficiency and the different phases of sleep. We also find different metrics to quantify one's lifestyle.

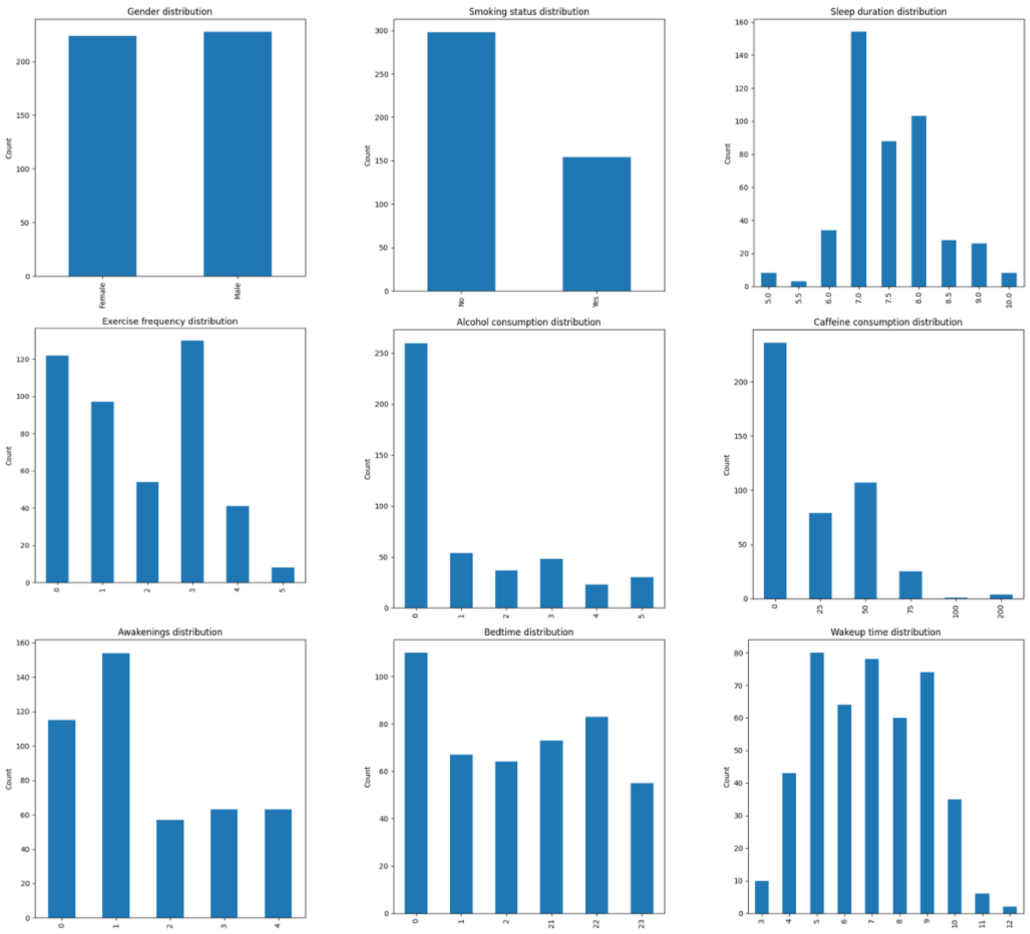
The dataset provided is very clean so there is not much pre-processing needed. We change the Bedtime and Wakeup time features to only reflect the hour that one goes to sleep and when one wakes up to make this feature more meaningful. We also change certain integer features to be categorical instead of numerical and proceed with exploring the distributions of individual variables.

Continuous Variables:



- Age appears to be fairly uniform with slight left and right tails, indicating a modestly higher number of older individuals in the dataset.
- Sleep efficiency has a left-skewed distribution and this suggests that most individuals in the dataset experience relatively efficient sleep.
- REP sleep percentage shows a somewhat platykurtic distribution with heavy tails. This indicates while there is a common range where most REM percentages fall, there are also individuals with unusually high or low REM percentages.
- Deep sleep and light sleep percentages are inversely related to each other, by nature of the way its calculated. Both their distributions appear to bimodal and this suggests there are 2 distinct groups within the dataset which could indicate different sleep patterns or quality. Most individuals have a high deep sleep percentage (low light sleep percentage) with a few outliers having very low deep sleep percentage (high light sleep percentage).
- Daily steps is distributed somewhat uniformly. This suggest most individuals have a moderate number of daily steps.

Categorical Variables:

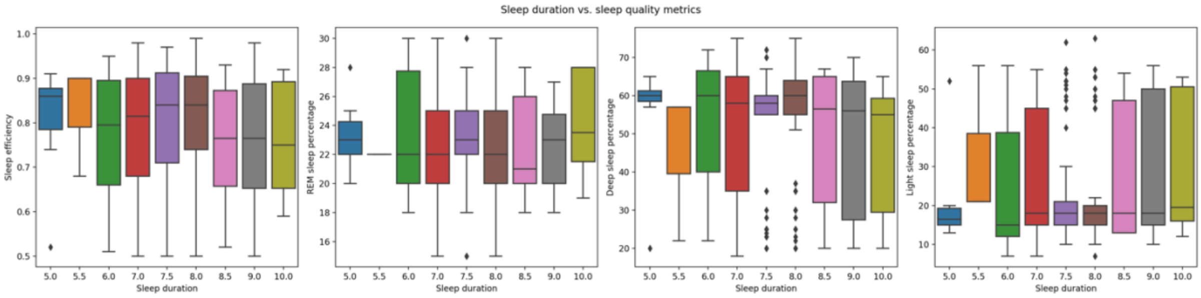


- Gender appears to be quite balanced, which is advantageous for comparative analysis as it reduces bias that could result from uneven sample
- A significantly large proportion of the sample are non-smokers
- Most common sleep dation ranges from 7-8 hours per night which indicates that most people are getting sufficient sleep based on scientific guidelines, with some outliers of the sample sleeping way less at 5 hours and way more at 10 hours.
- Most common frequency of exercise is 3 times a week, followed by no exercise. Very few people exercise 5 times a week.
- Most individuals do not consume alcohol before sleep, which could be due to the common advice of avoiding alcohol for better sleep quality.
- Similar to alcohol consumption, a large majority of individuals do not consume caffeine before sleep, indicating an awareness of its potential to disrupt sleep.
- Many individuals report few to no awakenings throughout the night, with the count decreasing as the number of awakenings increases. This suggests that frequent awakenings are less common among the participants.
- Most people go to sleep at or before 12am. The second most common time for people to go to sleep would be 10pm.
- Fewer people wake up very early or very late. Most individuals wake up around the middle of the morning, with decreasing counts towards the extremes.

Inter-variable relationships

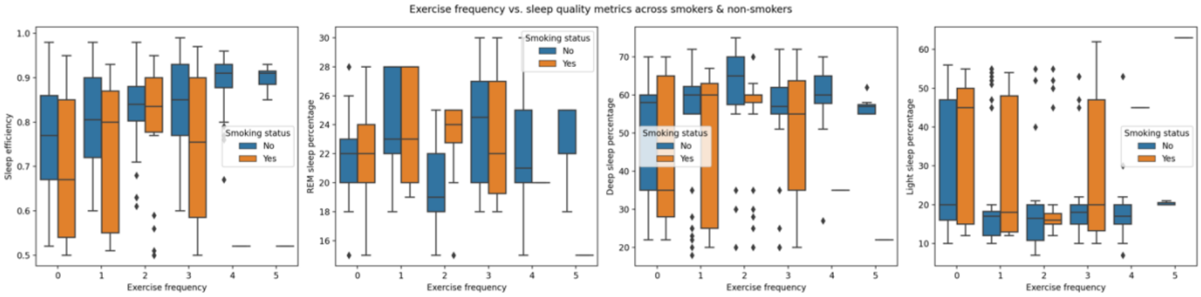
Here, we present some interesting findings from studying the relationships between our variables. We choose to quantify sleep quality with 4 main continuous metrics – Sleep efficiency, Deep sleep percentage, REM sleep percentage and Light sleep percentage.

1. **Sleep Duration vs. Sleep Quality**

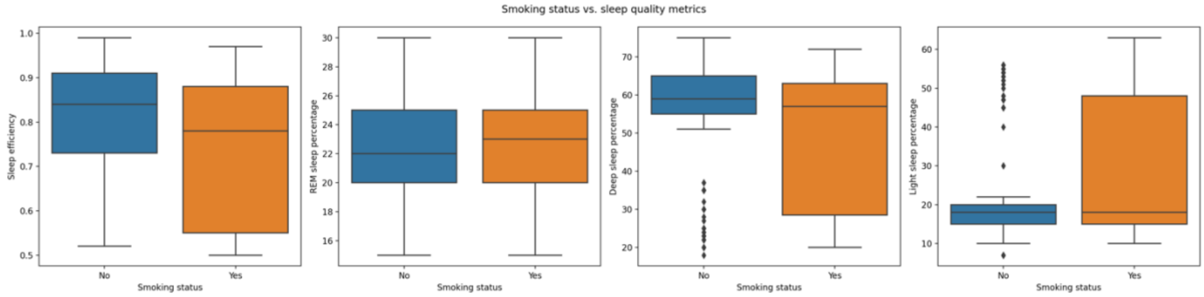


There does not appear to be a straightforward relationship between sleeping longer and improved sleep quality, at least in terms of efficiency and sleep phase composition. This challenges the common notion that sleeping more directly translates to better sleep quality. There seems to be a wide range of sleep efficiency values across all sleep durations, which suggests individual differences in sleep quality that are not solely dependent on the length of sleep. Some of the most efficient sleep is actually observed in the middle range of sleep duration rather than at higher durations.

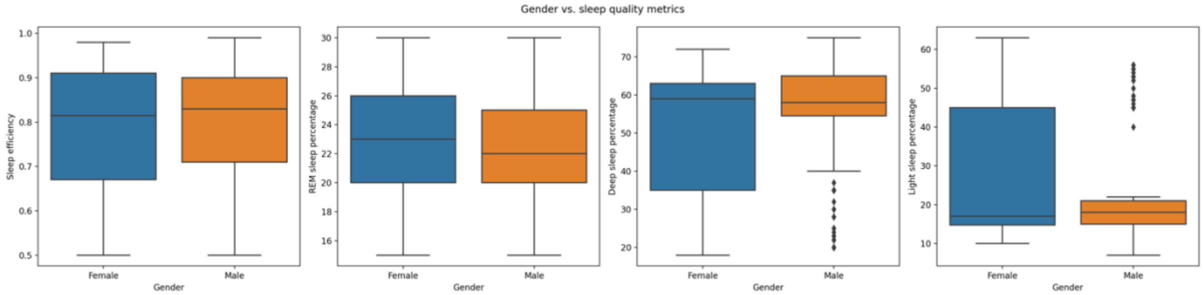
2. **Exercise and Smoking vs. Sleep Quality**



The interesting finding here is that there are instances where smokers who exercise frequently report better sleep quality compared to non-smokers who do not exercise at all. Looking at the box plots of sleep efficiency and deep sleep percentages, we observe that there is a chance for smokers who exercised just 3 times a week to have better sleep efficiency and deep sleep percentages. This could imply that the benefits of regular exercise might offset some negative impacts of smoking on sleep, although this does not necessarily suggest that smoking is less harmful overall since we see quite a consistent trend that smokers tend to have poorer sleep quality as seen below.



3. **Gender vs. Sleep Quality**



There is a very interesting comparison between male and female sleep quality metrics – the data seems to suggest that males, on average, have better sleep efficiency and greater percentages of deep sleep compared to females. Additionally, males also seem to have a lower percentage of light sleep. These differences may suggest that males appear to have better sleep quality and also raises questions about biological and lifestyle factors across genders that influence sleep.

Answer to Q2:

While is it important to not interpret the results above in silo, we propose 2 research questions given the interesting findings above:

1. How well can sleep duration, exercise frequency, smoking status, and gender predict one's sleep efficiency?
2. How well can one's lifestyle and health choices predict sleep quality?

Answer to Q3:

1. How well can sleep duration, exercise frequency, smoking status, and gender predict one's sleep efficiency?

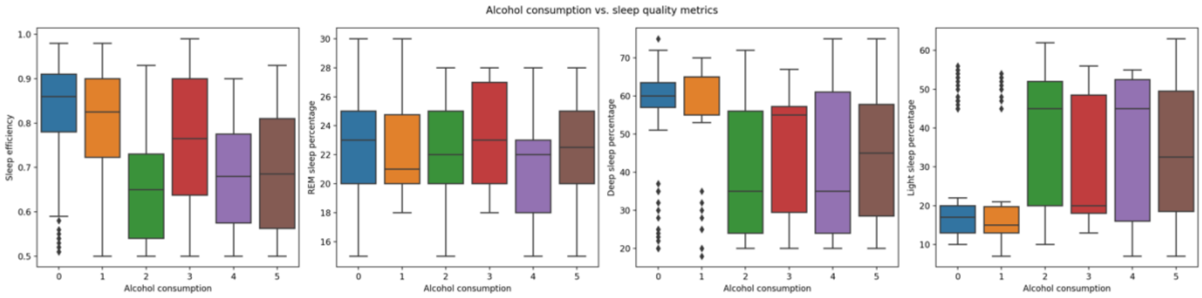
- Target variable: Sleep efficiency
- Input variables: Sleep duration, exercise frequency, smoking status and gender

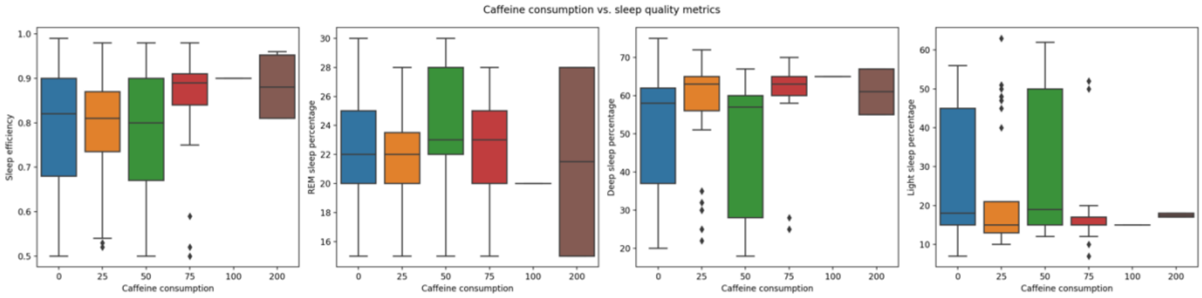
This question is chosen largely based on the interesting findings of inter-variable relationships above and we seek to understand the effects of each on sleep quality. Sleep efficiency is the main metric chosen to measure sleep quality. The proportion of time that is actually spent asleep can be more important than sleep duration as seen from the findings above. The input variables are chosen since we want to study how each feature influences the overall prediction of sleep efficiency.

2. How well can one's lifestyle and health choices predict sleep quality?

- Target variable: Deep sleep percentage
- Input variables: Caffeine consumption, Alcohol consumption, Smoking status, Exercise frequency, Daily steps

Alcohol and caffeine consumption does not show a very linear and strict relationship with sleep quality metrics.


Alcohol consumption vs. sleep quality metrics


Caffeine consumption vs. sleep quality metrics

Based on the interesting findings on exercise and smoking above, coupled with the nuanced impact of alcohol/caffeine consumption, we wish to isolate the effects of lifestyle choices and study how this impacts one's sleep quality. Deep sleep percentage is chosen to be the metric used to quantify sleep quality as it can be regarded as the most restorative portion of sleep.

Answer to Q4:

For both questions, a linear regression and decision tree model (CART) were used to model and predict Sleep efficiency and Deep sleep percentage respectively. The overall dataset is split into train and test splits using an 80-20 split. Results for the test set are presented below.

Predicting Sleep efficiency – Question 1

	MAE	MSE	RMSE	R Squared
Linear Regression	0.1092	0.0164	0.1282	0.1165
Decision Tree	0.1087	0.0164	0.1282	0.1174

The differences are very marginal and the 2 models are almost equivalent. However the decision tree model performs slightly better given its lower MAE and higher R squared, which indicates it fits the dataset better and is slightly more accurate in its predictions.

Predicting Deep sleep percentage – Question 2

	MAE	MSE	RMSE	R Squared
Linear Regression	10.3327	187.5164	13.6937	0.2322
Decision Tree	11.4287	220.6375	14.8539	0.0966

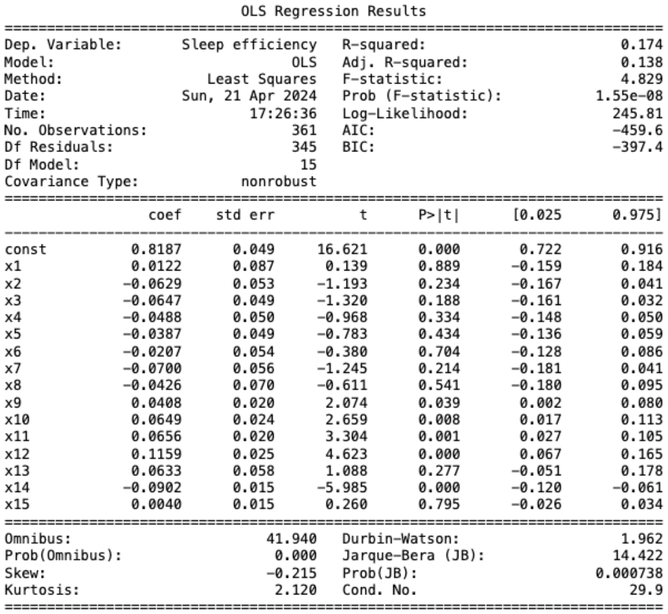
For predicting deep sleep percentage however, linear regression significantly outperforms decision tree as it has lower residual errors metrics and a higher R squared, which implies that it is capable of explaining a higher percentage of the dataset's inherent variance.

However, it is also worthwhile to note that the target variable deep sleep percentage is bimodal and not normally distributed. There also exists a lot of non-linear patterns between features and target. Furthermore, we also have not fully utilitized all available features to predict our target variable and have not done any hyperparameter tuning. This could therefore explain why our set of results is so poor.

Answer to Q5:

1. How well can sleep duration, exercise frequency, smoking status, and gender predict one's sleep efficiency?

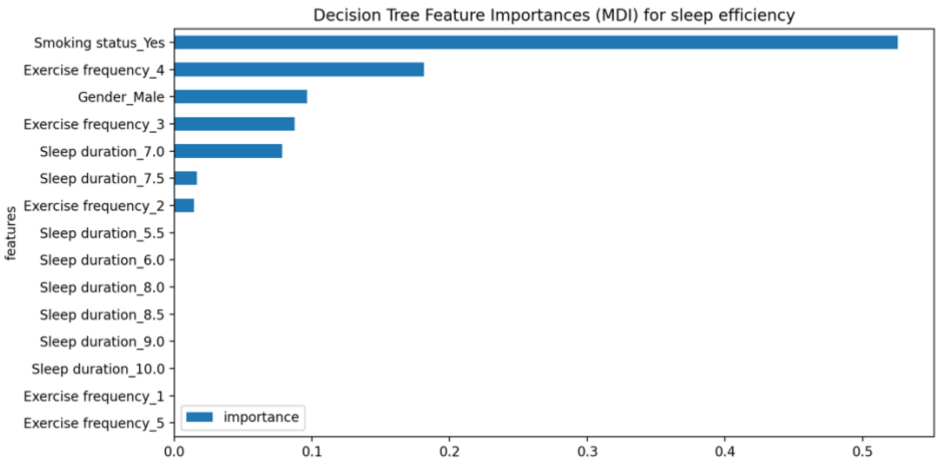
Sleep duration, exercise frequency, smoking status and gender can predict one's sleep efficiency to a certain extent. While the results from our prediction models above were not close to ideal, we can see from the linear regression coefficients table below that whether or not somebody exercises or smokes to be the greatest determinant in one's sleep efficiency.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:       Sleep efficiency   R-squared:                       0.174
Model:                            OLS   Adj. R-squared:                  0.138
Method:                 Least Squares   F-statistic:                     4.829
Date:                Sun, 21 Apr 2024   Prob (F-statistic):           1.55e-08
Time:                        17:26:36   Log-Likelihood:                 245.81
No. Observations:                 361   AIC:                            -459.6
Df Residuals:                     345   BIC:                            -397.4
Df Model:                          15                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.8187      0.049     16.621      0.000       0.722       0.916
x1             0.0122      0.087      0.139      0.889      -0.159       0.184
x2            -0.0629      0.053     -1.193      0.234      -0.167       0.041
x3            -0.0647      0.049     -1.320      0.188      -0.161       0.032
x4            -0.0488      0.050     -0.968      0.334      -0.148       0.050
x5            -0.0387      0.049     -0.783      0.434      -0.136       0.059
x6            -0.0207      0.054     -0.380      0.704      -0.128       0.086
x7            -0.0700      0.056     -1.245      0.214      -0.181       0.041
x8            -0.0426      0.070     -0.611      0.541      -0.180       0.095
x9             0.0408      0.020      2.074      0.039       0.002       0.080
x10            0.0649      0.024      2.659      0.008       0.017       0.113
x11            0.0656      0.020      3.304      0.001       0.027       0.105
x12            0.1159      0.025      4.623      0.000       0.067       0.165
x13            0.0633      0.058      1.088      0.277      -0.051       0.178
x14           -0.0902      0.015     -5.985      0.000      -0.120      -0.061
x15            0.0040      0.015      0.260      0.795      -0.026       0.034
==============================================================================
Omnibus:                       41.940   Durbin-Watson:                   1.962
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               14.422
Skew:                          -0.215   Prob(JB):                     0.000738
Kurtosis:                       2.120   Cond. No.                         29.9
==============================================================================
```

Legend
x1: Sleep duration_5.5,
x2: Sleep duration_6.0,
x3: Sleep duration_7.0
x4: Sleep duration_7.5
x5: Sleep duration_8.0
x6: Sleep duration_8.5
x7: Sleep duration_9.0
x8: Sleep duration_10.0
x9: Exercise frequency_1
x10: Exercise frequency_2
x11: Exercise frequency_3
x12: Exercise frequency_4
x13: Exercise frequency_5
x14: Smoking status_Yes
x15: Gender_Male

Note that the reason a lot more variables are present in our linear regression is due to the fact that we have one-hot encoded them and pre-processed them. We can see from the OLS regression results that variables x9 to x12 and x14 are statistically significant predictors (p-value < 0.05) for sleep efficiency. These correspond to the one-hot encoded variables for exercise and smoking status. The positive sign of the coefficients indicate the positive relationship between exercising and sleep efficiency. Observe that the more one exercises, the higher the coefficient. This indicates that exercising more frequently can have positive impacts to sleep efficiency. On the other hand, the negative coefficient corresponding to x14, suggests that smoking is not beneficial to one's sleep efficiency.

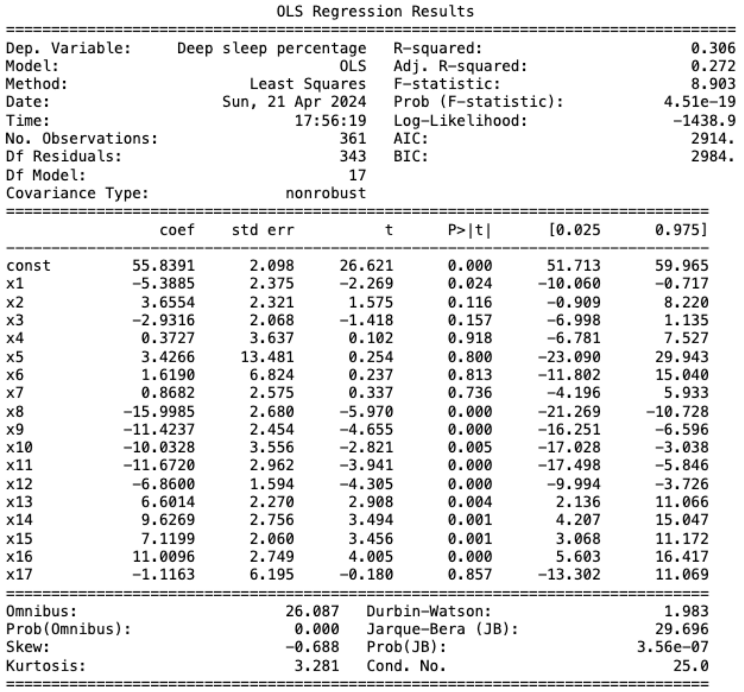


Likewise, from the decision tree regressor, we output a feature importances plot shown above and find that smoking and exercise are the 2 most important determinants to one's sleep efficiency, which corroborates our findings from the linear regression model.

It is also interesting to note that one's sleep duration has close to no impact in predicting sleep efficiency which is in agreement to the data exploratory box plots above. Gender also does not seem to play a significant role, this implies that the findings we drawn above might be due to exogenous factors unrelated to gender, but have an underlying correlation with males/females – for e.g, it may be the case that females exercise less, leading to a poorer sleep efficiency score.

2. How well can one's lifestyle and health choices predict sleep quality?

While one's lifestyle and health choices can influence and predict sleep efficiency to a certain extent, we now turn to predicting sleep quality.



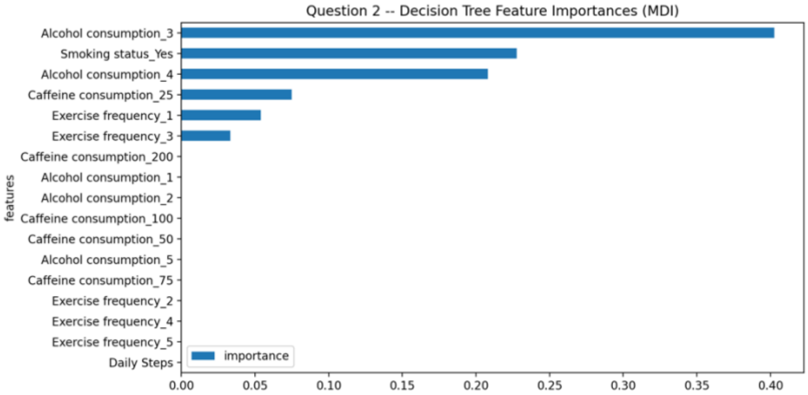
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Deep sleep percentage   R-squared:                       0.306
Model:                               OLS   Adj. R-squared:                  0.272
Method:                    Least Squares   F-statistic:                     8.903
Date:                   Sun, 21 Apr 2024   Prob (F-statistic):           4.51e-19
Time:                           17:56:19   Log-Likelihood:                -1438.9
No. Observations:                    361   AIC:                             2914.
Df Residuals:                        343   BIC:                             2984.
Df Model:                             17
Covariance Type:               nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         55.8391      2.098     26.621      0.000      51.713      59.965
x1            -5.3885      2.375     -2.269      0.024     -10.060      -0.717
x2             3.6554      2.321      1.575      0.116      -0.909       8.220
x3            -2.9316      2.068     -1.418      0.157      -6.998       1.135
x4             0.3727      3.637      0.102      0.918      -6.781       7.527
x5             3.4266     13.481      0.254      0.800     -23.090      29.943
x6             1.6190      6.824      0.237      0.813     -11.802      15.040
x7             0.8682      2.575      0.337      0.736      -4.196       5.933
x8           -15.9985      2.680     -5.970      0.000     -21.269     -10.728
x9           -11.4237      2.454     -4.655      0.000     -16.251      -6.596
x10          -10.0328      3.556     -2.821      0.005     -17.028      -3.038
x11          -11.6720      2.962     -3.941      0.000     -17.498      -5.846
x12           -6.8600      1.594     -4.305      0.000      -9.994      -3.726
x13            6.6014      2.270      2.908      0.004       2.136      11.066
x14            9.6269      2.756      3.494      0.001       4.207      15.047
x15            7.1199      2.060      3.456      0.001       3.068      11.172
x16           11.0096      2.749      4.005      0.000       5.603      16.417
x17           -1.1163      6.195     -0.180      0.857     -13.302      11.069
==============================================================================
Omnibus:                       26.087   Durbin-Watson:                   1.983
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               29.696
Skew:                          -0.688   Prob(JB):                     3.56e-07
Kurtosis:                       3.281   Cond. No.                         25.0
==============================================================================
```

Legend

x1: Caffeine consumption_25
x2: Caffeine consumption_50
x3: Caffeine consumption_75
x4: Caffeine consumption_100
x5: Caffeine consumption_200
x6: Alcohol consumption_1
x7: Alcohol consumption_2
x8: Alcohol consumption_3
x9: Alcohol consumption_4
x10: Alcohol consumption_5
x11: Smoking status_Yes
x12: Exercise frequency_1
x13: Exercise frequency_2
x14: Exercise frequency_3
x15: Exercise frequency_4
x16: Exercise frequency_5
x17: Daily Steps

We see that variables x8-x16 are all statistically significant features ant they correspond to alcohol consumption before bed, smoking and exercise frequency. This is in line with the findings from answering question 1. It is also noteworthy to point out that the coefficients for the one-hot encoded alcohol consumption variables are highly negative and their magnitudes are very large compared to other features. This implies that high alcohol consumption between bed can significantly impact deep sleep percentage. Moderate alcohol consumption may still be tolerated. Smoking and more frequent exercise are also 2 main determinants of deep sleep percentage.

Although x1 is a statistically significant predictor, by and large, caffeine consumption before bed is not statistically significant when predicting deep sleep percentage.



Looking at the decision tree's feature importances plot for predicting deep sleep percentage also corroborates our findings from looking at the coefficients and p-values of our linear regression model. This overall suggests that one's lifestyle and health choices, particularly alcohol consumption before bed, smoking and exercising can greatly influence one's sleep quality.

Answer to Q6:

The analysis of sleep data using Linear Regression and Decision Tree models has revealed several key findings:

- Sleep Efficiency: Predicted moderately well by both models, with the Decision Tree model showing a slightly better fit.
- Deep Sleep Percentage: Strongly predicted by the Linear Regression model, outperforming the Decision Tree significantly.

While both models fail to predict sleep efficiency and deep sleep percentage to a large extend, the regression summary tables and decision tree feature importances score highlight the importance of frequent exercise, not smoking, and not consuming alcohol before bed for a better sleep efficiency and deep sleep percentage.

Answer to Q7:

These models can be used in health and wellness applications to provide personalized sleep improvement recommendations. If it were to be designed for a real-world application, more steps have to be taken to ensure the model's robustness against inherent biases and overfitting. This would mean hyperparameter optimizing and perhaps a more complex model might have to be used, building on CART, such as ensemble models.

More features would have to be used on top of the ones that we've already discussed. This would mean collecting more nuanced data from users' and their lifestyle and health choices, possibly their current weight, diet, proxy indicators for stress levels etc. The model could predict sleep quality and efficiency and then suggest changes to optimize sleep. For instance, a wellness app could use these models to alert users if their current habits may be adversely affecting their sleep, and offer tailored advice for improvement.